Machine Learning for Astrophysics

We are witnessing an exponential increase in size of the data as different instruments (e.g. telescopes) are continuously observing the sky and collecting data of huge number of objects in universe. Only in the last decade, hundreds of terabytes of data on millions of sources have been collected (after in situ data reduction), and petabyte data of billions of objects are predicted to be collected during the next decade. To get a sense of the rapid increase in astronomical data, take the upcoming Large Synoptic Survey Telescope as an example. LSST takes 2000 nightly images, each several GBs in size. Of these images, 10 million sources are processed and reduced to 100 thousand transient events that results in more than 10 terabytes of data, every night.

With the traditional manual way of analyzing and understanding astronomical data, we are far behind in converting them to knowledge. Machine learning can help us effectively and efficiently analyze this massive size of data, by learning what domain experts know (e.g. classification of objects) or learning to report what they don't know (e.g. finding rare/never-seen events). Example applications are discovering new planets from transit signals; morphological classification of galaxies; classifications of objects to galaxy, stars, etc.; photometric redshift, and efficient search of transient astronomical events. Most of these applications are big data **science** problems where there are a huge number of instances to learn from in order to increase our knowledge of the universe.

Meanwhile, to collect this data we are making new instruments, the components of which are built for the first time, e.g. Fine Guidance Sensors (FGS) in the Hubble Space telescope, and the Kepler space telescope. Given that the domain knowledge is very limited for these complex units that are built specifically for a given instrument, the hand-coded physics-based prognostics is not accessible. One approach is to explore the data-driven methodologies to analyze the sensory data of the unit and understand how the states of a unit are related to its degradation over time. The challenge in these **engineering** problems is that there are usually only a few working and failed units available to learn from, an extreme small sample size problem.

Whether it is a science problem or an engineering one, specialized machine learning algorithms need to be developed to overcome the unique challenges astronomical data introduce. Some of these challenges are: dealing with sample selection bias (e.g. due to data reduction in the instrument or acquisition constraints), inconsistent data representation from different instruments, learning from small sample size problem (e.g. in engineering problems where only a few instances are available), and lack of interpretability (compared to physical modeling).

In this session, we will discuss these and other challenges.